A Semi-linear Model for Normalization and Analysis of cDNA Microarray Data

Jian Huang[1,2*], *Hsun-Chih Kuo*[2], *Irina Koroleva*[3], *Cun-Hui Zhang*[4], *and Marcelo Bento Soares*[3,5]

[1]*Departments of Statistics and Actuarial Science,* [2]*Biostatistics Division of Statistical Genetics,* [3]*Pediatrics,* [5]*Biochemistry, Orthopaedics, Physiology and Biophysics, The University of Iowa, Iowa City, IA 52242, USA and* [4]*Department of Statistics, Rutgers University, Piscataway, NJ 08855, USA*

ABSTRACT

Motivation: Microarray analysis is a technology for monitoring gene expression levels on a large scale and has been widely used in functional genomics. A challenging issue in the analysis of microarray data is normalization. A proper normalization procedure ensures that the intensity ratios provide meaningful measures of relative expression levels. There are two important questions concerning normalization not adequately addressed in the current literature: (a) how to identify genes that have constant expression levels in order to establish the normalization curves; (b) how to account for the uncertainty inherent in the normalization process in the subsequent statistical analysis.

Results: We propose a semi-linear model that incorporates normalization into the analysis. This method does not make the usual assumptions needed for the loess and dye-swap normalization procedures, nor does it require to identify a set of constantly expressed genes prior to normalization. It also naturally accounts for the uncertainty in the normalization process. We apply the proposed method to two microarray data sets to illustrate this approach and its differences from the loess normalization method.

Availability: A set of programs will be electronically sent upon request.

Contact: jian-huang@uiowa.edu

1 Introduction

The cDNA microarray technology is a powerful tool for monitoring gene expression levels on a large scale and has been widely used in functional genomics (Brown and Botstein, 1999). An important issue in analyzing microarray data is normalization. The purpose of normalization is to ensure that the intensity levels from the two florescent dyes are comparable (Yang et al. 2000).

In a microarray experiment, many factors contribute to the possible bias and variation of the primary data output — the hybridization intensities with the reporters of the genes in the bio-samples. These factors include differential efficiency of dye incorporation, differences in concentration of DNA on arrays, batch bias, difference in the amount of RNA labeled between the two channels, experimental variability in DNA extraction and reverse transcription, variability in probe coupling and washing process, unevenness of the slide surface, differences in the printing pin heads, and so on. Therefore, proper normalization is a critical component in the analysis of microarray data.

Ideally, normalization should be done using genes whose expression levels remain constant and cover the whole dynamic range of the intensity. However, in a microarray experiment, it is usually not known a priori which genes exhibit differential expression levels and which genes do not. Indeed, the purpose of many experiments is to identify differentially expressed genes.

Under the assumption that there is only a small percentage of the genes in the study that will have differential expressions, Yang et al. (2000) proposed fitting a local regression (loess) curve (Cleveland 1979) for normalization using all the genes. The rationale is that if the number of differentially expressed genes is relatively small, then the loess normalization curve should not be affected significantly by the differentially expressed genes. Thus this approach should work well if the investigator knew that there are only a small number (and a small percentage) of genes that have differential expressions. However, there are also cases in which the investigator may not know how many genes have differential expressions, and it may not be reasonable to assume that the percentage of differentially expressed genes is small. A further question is that it is not easy to quantify how small is small in order to ensure that the loess normalization using all the genes will not in itself introduce bias.

If it is expected that many genes will have differential expressions, Yang et al. suggested using dye-swap for normalization. This approach makes the assumption that the normalization curves in the two dye-swaped slides are symmetric about the horizontal axis in an M-A plot. Because of the slide-to-slide variation, this assumption may not always be satisfied.

In Tseng et al. (2001), they first used a rank based procedure to select genes that are likely to be non-differentially expressed, and use these genes in loess normalization. Although selection of genes based on ranks

*To whom correspondence should be addressed

is more robust than using the actual values, there is potential bias in the selection process because ranking based on un-normalized data may not be correct. There is also uncertainty associated with the rank selection procedure. In addition, a threshold value is required in this rank based procedure. How sensitive the final results depending upon the threshold value may need to be evaluated on a case by case basis.

In both Yang et al. (2001) and Tseng et al. (2001), normalization is considered as a step separated from the subsequent statistical analysis, and as such, is dealt with separately. We propose a method that treats normalization as an integrated part of the overall analysis, in the same spirit as in the analysis of covariance (ANCOVA). In ANCOVA, the confounding factors are adjusted for in assessing the "treatment effect" in a single regression model. In a cDNA microarray experiment, the contribution of the different dye efficiency and other experimental factors to the observed difference in the intensity levels can be considered as confounding factors. The difference in intensity levels that is biologically meaningful can be considered as the "treatment effect" in the ANCOVA terminology. Thus in our proposed approach, normalization is a component of the model that adjusts for the confounding factors, and the true biological difference in gene expressions is the component of primary interest. Both components are estimated simultaneously. In this way, there is no need to know which genes or how many genes have constant expressions á priori.

More specifically, we propose using a semi-linear (SL) model for simultaneous intensity-dependent normalization and analysis of gene expression data. The original form of the SL model was first proposed by Engle et al.(1986) in a study of relationship between weather and electricity sales, while adjusting for other factors. In the original form of the SL model, there is only one nonparametric component. We extend this model so that it is suitable for microarray data. In this SL model for microarray data, normalization is considered as a nonparametric component, and normalization is done within the analysis, not as a separate step.

Below, we first describe the SL model for microarray data. In Section 3, we describe an algorithm for computing the normalization curves and the estimated expression levels based on the SL model. We also consider a bootstrap approach for making statistical inference for the SL model in order to identify differentially expressed genes. In Section 4, we illustrate the proposed method by two examples. Some concluding remarks are given in Section 5.

2 A Semi-linear Model for Microarray Data

We now describe a SL model for microarray data in some generality. This model can be applied to both the reference design and comparative design by suitably coded covariates.

Let J be the number of genes (or ESTs) in the study. Suppose that there are n slides in the experiment. Let R_{ij} and G_{ij} be the red (Cy 5) and green (Cy 3) intensities of gene j in slide i, respectively. Let y_{ij} be the log-intensity ratio of the red over green channels of the jth gene in the ith slide, and let x_{ij} be the corresponding average of the log-intensities of the red and green channels. That is,

$$y_{ij} = \log_2 \frac{R_{ij}}{G_{ij}}, \;\; x_{ij} = \frac{1}{2}\log_2(R_{ij}G_{ij}),$$

$i = 1, \ldots, n, j = 1, \ldots, J$. Let z_i be a covariate vector associated with the ith slide. It describes the characteristics of the ith slide, and can also be used to code various types of designs. The SL model is

$$y_{ij} = \phi_i(x_{ij}) + z_i'\beta_j + \varepsilon_{ij}, \tag{1}$$

$i = 1, \ldots, n, j = 1, \ldots, J$, where β_j is the effect associated with the jth gene; ε_{ij} is the error term with mean zero and unknown variance σ_{ij}^2; and the function ϕ_i is a nonparametric component of the model and is to be estimated based on the data.

The function ϕ_i is the normalization curve for the ith slide, because it is the difference in the log intensities of red and green channels, given the total log intensities, in the absence of the gene effects. Therefore, ϕ_i represents the baseline log intensity ratios of the genes with constant expressions. We note that in model (1), it is only made explicit that the normalization curve ϕ_i is slide-dependent. It can also be made to be dependent upon regions of the slides to account for spatial effect. For example, it is straightforward to extend the model with an additional subscript in (y_{ij}, x_{ij}) and ϕ_i and make ϕ_i also depend on the printing-pin blocks within a slide.

As mentioned above, we can code various types of designs by use of indicator variables as covariate z_i (Drapper and Smith 1980). For example, in a reference design that compares diseased and normal tissues using a common reference (Kerr and Churchill 2001), we can let z_i take the value $(1, 0)$ for the diseased tissue and $(0, 1)$ for the normal tissue. For a multiple k-sample comparison problem, we can use a k dimensional indicator covariate. We note here that we did not explicitly include an intercept term in the model. If an intercept is included, then only a $(k-1)$-dimensional indicator vector is needed for a k-sample comparison.

In a loop design (Kerr and Curchill 2001), in which multiple tissues are compared directly without using a common reference, we can also use indicator variables to code the experiment. In the simplest case where two samples are hybridized on the same slide, the SL model (1) becomes

$$y_{ij} = \phi_i(x_{ij}) + \beta_j + \varepsilon_{ij}, \tag{2}$$

where β_j represents the difference in the expression levels of gene j after normalization.

3 Estimation and inference in SL Model

We use a semi-parametric approach for estimating $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ simultaneously. Many smoothing procedures can be used here for the estimation of $\boldsymbol{\phi}$. We choose the method of polynomial spline (Schumaker 1981). This method is easy to implement, and has similar performance as other nonparametric curve estimation approaches such as loess (Friedman et al. 2001).

Let $b_1, \ldots, b_Q$ be Q B-spline base functions. We approximate ϕ_i by

$$\lambda_{s0} + \sum_{q=1}^{Q} b_q(x)\lambda_{iq} \equiv \boldsymbol{b}(x)'\lambda_i,$$

where $\boldsymbol{b}(x) = (1, b_1(x), \ldots, b_Q(x))'$, and $\lambda_i = (\lambda_{i0}, \lambda_{i1}, \ldots, \lambda_{iQ})'$ are coefficients to be estimated from the data. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. It is possible to use a data-driven approach to determine Q for each i, so that each ϕ_i uses a different number of spline basis functions. However, here we choose to use the same number of spline basis functions for each normalization curve. This has the advantage of keeping the normalization consistent across the slides.

We use a least squares (LS) criterion in our estimation. Let $\boldsymbol{w} = \{w_{ij}\}$ be a matrix of user specified weights. The simplest choice is to let $w_{ij} = 1$. We incorporate this weight matrix into the objective function to allow for incorporation of quality measurements of the spots into the analysis. For instance, we can let w_{ij} be reciprocal to the standard deviations associated with the log-intensity ratios. Such standard deviations can be computed from the standard deviations from pixel intensities which are usually available from microarray data output files. We can also use $\boldsymbol{w}$ as a filter to screen out or down weighting lesser quality spots. The weighted LS criterion function is

$$D_w(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^2[y_{ij} - \boldsymbol{b}(x_{ij})'\lambda_i - z_i'\beta_j]^2. \quad (3)$$

Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$. One approach for minimizing $D_w(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is to use the Gauss-Seidel method, also called back-fitting algorithm (Hastie et al. 2001), that alternately updates $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$. Set $\boldsymbol{\lambda}^{(0)} = 0$. For $k = 0, 1, 2, \ldots$,

Step 1: Compute $\boldsymbol{\beta}^{(k)}$ by minimizing $D_w(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

Step 2: For the $\boldsymbol{\beta}^{(k)}$ computed above, obtain $\boldsymbol{\lambda}^{(k+1)}$ by minimizing $D_w(\boldsymbol{\lambda}, \boldsymbol{\beta}^{(k)})$ with respect to $\boldsymbol{\lambda}$.

Iterate Steps 1 and 2 until the desired convergence criterion is satisfied. Because the objective function is strictly convex, the algorithm converges to the unique global optimal point. Suppose that the algorithm converges at step K. Then the estimated values of β_j are $\widehat{\beta}_j = \beta_j^{(K)}, j = 1, \ldots, J$, and the estimated normalization curves are

$$\widehat{\phi}_i(x) = \boldsymbol{b}(x)'\lambda_i^{(K)},\ i = 1, \ldots, n.$$

For the two-sample comparison problem described by model (2), and suppose the weights are all 1, Step 1 above becomes simply taking the averages:

$$\beta_j^{(k)} = n^{-1}\sum_{i=1}^{n}[y_{ij} - \boldsymbol{b}(x_{ij})'\lambda_i^{(k)}],\ j = 1, \ldots, J.$$

The algorithm described above can be conveniently implemented in the statistical computing environment R (Ihaka and Gentleman, 1996). Specifically, Steps 1 and 2 are least squares problems, which can be solved by the function lm in R. The function bs can be used to create a basis matrix for the polynomial splines.

After obtaining the estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$, it is also desirable to estimate the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$. In principle, it can be estimated in a similar way as in the linear model. However, computation of this variance-covariance matrix involves inverting a $G \times G$ matrix. When $G \approx 10^4$, as in many microarray experiments, direct inversion of this matrix can be difficult. One approach to get around this difficulty is to use the bootstrap (Efron 1979; Efron and Tibshirani 1993). Although this is also computationally intensive, it does not require inverting a high dimensional matrix. We propose using the following bootstrap scheme. Let B be the bootstrap replication size.

(1) For slide $i, i = 1, \ldots, n$, resample with replacement from $(y_{ij}, x_{ij}), j = 1, \ldots, J$ to obtain the bth bootstrap sample $(y_{ij}^{(b)}, x_{ij}^{(b)}), j = 1, \ldots, J$.

(2) Fit the semilinear model using the bth bootstrap sample $(y_{ij}^{(b)}, x_{ij}^{(b)}, z_i), j = 1, \ldots, J; i = 1, \ldots, n$ to obtain the bootstrap estimate $\boldsymbol{\beta}^{(b)}$. In the above, we have kept the covariate z_i fixed.

(3) Compute the sample variance of the bootstrap estimates $\boldsymbol{\beta}^{(b)}, b = 1, \ldots, B$. This bootstrap sample variance is used as an estimation of the variance of $\widehat{\boldsymbol{\beta}}$.

There are more than one way to bootstrap in the current setting. For example, we can also resample the residuals after fitting the SL model. We use the bootstrap scheme described above because its resampling process most faithfully keeps the original data structure. Also, bootstrapping both y_{ij} and x_{ij} relies less on the model assumption (Efron and Tibshirani, 1993).

Instead of the LS estimate, we can also use robust regression approaches. For example, we can replace the L_2 norm in the objective function by the L_1 norm, which would result in the least absolute deviation (LAD) regression. We can also use other robust regression objective functions, such as Tukey's biweight function. The above iterative computation steps can still be applied with the chosen objective function. However, computation in each step will be more demanding.

4 Examples

We now illustrate the SL model for microarray data by two data sets. As a point of comparison, we also consider the loess normalization method of Yang et al. (2000). This is mainly to illustrate the differences between the proposed normalization method and the loess normalization. We focus on the normalization and its effect on the subsequent analysis. Therefore, we will not discuss some important issues, such as the problem of multiple comparisons in determining the significance of the findings.

4.1 Apo AI data

The purpose of this experiment is to identify differentially expressed genes in the livers of mice with very low HDL cholesterol levels compared to inbred mice (Callow et al. 2000). The treatment group consists of 8 mice with the apo AI gene knocked-out and the control group consists of 8 C57B1/6 mice. For each of these mice, target cDNA is obtained from mRNA by reverse transcription and labeled using a red fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations was obtained by pooling cDNA from the 8 control mice. The target cDNA is hybridized to microarrays containing 5,548 cDNA probes.

This data set was analyzed by Callow et al. (2000) and Dudoit et al. (2000). Their analysis consists of the following steps: (1) loess normalization; (2) computation of a two sample t-statistic for each gene; (3) permutation to estimate the distribution the t-statistics and; (4) using the Westfall and Young step-down method for adjusting p-values to correct for multiple comparison (Westfall and Young 1993). Eight genes with adjusted p-value ≤ 0.01 are identified and subsequently sequence verified. These genes are listed in the left panel of Table 1, in the order of the magnitude of their t-statistics.

We apply the proposed normalization and analysis method to this data set. As in Dudoit et al., we use printing-tip dependent normalization. The SL model used here is

$$y_{ijp} = \phi_{ip}(x_{ijp}) + z_i'\beta_j + \varepsilon_{ijp},$$

where $i = 1, \ldots, 16$, $j = 1, \ldots, \#$ of spots in each block, $p = 1, \ldots, 16$. Here j and i index genes and slides as before, p indexes the pth printing-tip block. ε_{ijp}'s are assumed to be i.i.d. normal with mean 0 and variance σ^2. The covariate $x_s = (1, 0)'$ for the treatment group (apo AI knock out mice) and $x_s = (0, 1)'$ for the control group (C57B1/6 mice). The coefficient $\beta_j = (\beta_{j1}, \beta_{j2})$. The contrast $\beta_{j1} - \beta_{j2}$ measures the expression difference for the jth gene between the two groups. The variances of the estimators of this contrast are obtained by the bootstrap with bootstrap sample size $B = 500$ described in Section 3.

As an example of the normalization results, Figure 1 displays the M vs. A plot and the printing-tip dependent normalization curves in blocks 1, 4, 5, 9, 12, and 16 for the data from knock-out mouse 1. The green line is the normalization curve based on the SL model, and the dotted red line is the loess normalization curve. The degree of freedom used in the spline basis function in the SL normalization is 12, and the span used in the loess normalization is 0.40. We see that, although the overall trend of the two normalization curves are similar, there are indeed differences between the two normalization curves from the two methods.

Tables 1 and 2 lists the top 16 genes identified based on the proposed method and the method of Dudoit et al. (2000), respectively. In Table 1, the numerator is the estimated expression difference between the two groups based on the SL model, the denominator is the standard error computed based on the bootstrap. The t-statistic is the ratio of the two. The p-values are given as a point of reference of the "significance" of the t-statistics, they are approximate and should not be interpreted in a rigid fashion. As mentioned earlier, we are mainly interested in the comparison of methods, so no multiple comparison adjustment is made to these p-values. This does not affect our comparison of the different methods. These remarks also apply to the second example below.

In both tables, the genes are listed in the order of the magnitude of their t-statistics. Eight of the 16 genes (ID 540, 2149, 5356, 4941, 4139, 1496, 2537, 1739) in Table 1 are the same as those identified by Dudoit et al. However, the orders of the genes in the lists are different. It is also noticeable that many genes in these two lists are different. For example, gene 541 is ranked number 7 in our analysis, but it is not in the top 16 list based on the method of Dudoit et al.(2000).

Figure 3(a) shows the scatter plot of the estimated mean expression differences based on the SL versus those calculated after the loess normalization. The colored spots represent the top 16 genes identified from both methods. The yellow spots are the 8 genes common from both methods. The red spots are the remaining 8 ones from the SL method, and the green ones are the remaining 8 from the loess method. Figure 3(b) shows the boxplots of the expression differences based on the two methods. We see from these two plots that the bulk of the expression differences have the same distribution. However, at the high end of expression differences, the values from the SL model tend to be higher. The correlation coefficient between the two expression differences is 0.85.

There are two reasons that the results are different. First, the different normalization procedures give different expression differences. For example, for gene 2149, the mean difference based on the loess normalization is 3.0806, the mean difference based the proposed method is 3.3294. For the eight genes identified by Dudoit et al., the proposed method yields bigger mean differences. Because normalization is done separately and use all the genes, the differentially expressed genes tend to pull the normalization curve towards themselves. Therefore, it may

lead to underestimates of the mean differences.

Second, the estimates of the standard errors are also different. The estimates based on the individual genes have a relatively large range, from 0.016 to 0.53. As we can see from Table 2, the standard error estimates based on the individual genes are quite erratic. The estimates based on the SL model ranges from 0.19 to 0.23, and is centered around 0.21. We again note that these estimated standard errors based on the SL model are adjusted for the normalization, and depend on the total log-intensities.

4.2 Hippocampus data

The hippocampus experiment was carried out in the Functional Genomics Laboratory at the University of Iowa College of Medicine. RNA samples were collected from mouse hippocampus and from the remainder of the brain. 557 clones originating from non-normalized, normalized cDNA libraries (Bonaldo et al. 1996) and from a cDNA library of rare mRNA messages from mouse hippocampus were selected. The original purpose of this experiment was to compare the results of expression levels based on sequence analysis with those based on the microarray analysis. However, this experiment also provides us an opportunity to test our proposed method.

cDNA samples of tissues from hippocampus and the remainder of the brain in mice were prepared and hybridized to 12 printed slides. Each slide contains 557 arrayed mice DNA, each of these 557 genes were printed 4 times on each slide. Some of the clones were not verified in the sequence analysis. However, they are used in the normalization methods, but are not considered in the discussion of the results below. In 6 of these 12 slides, cDNA from brain (without hippocampus) were labelled with florescent Cy5, and cDNA from hippocampus were labelled with florescent Cy3. In the remaining 6 slides, the dye scheme is reversed.

For each clone, we computed a weighted average from the 4 printed spots as the intensity level for each channel. The weights are reciprocal to the variances associated with the spots. These weights are chosen to achieve the minimum variance among all the weighted averages. The variance associated with a spot in a slide is computed from the standard deviation of the pixels that constitute the spot and the number of the pixels. They are part of the data output from the GenePix 4000B scanner and the GenePix Pro software (Axon Instruments Inc.). The log intensity ratio y_{ij} and log intensity product x_{ij} is computed based on the weighted averages from the red and green channels for the jth gene in the ith slide , $i = 1, \ldots, 12; j = 1, \ldots, 557$. Here in the intensity ratio, the intensity level of hippocampus is always in the denominator, and the intensity level of brain without hippocampus is in the numerator, for the data from all the 12 slides. Because two tissue samples are compared directly without resorting to a reference sample, the SL model used for this data set is simply

$$y_{ij} = \phi_i(x_{ij}) + \beta_j + \varepsilon_{ij}, i = 1, \ldots, 12,\ j = 1, \ldots, 557;$$

where β_j measures the difference in expression levels of the jth gene after normalization. Here we assume ε_{ij}'s are i.i.d. $N(0, \sigma^2)$.

Figure 2 shows the M-A plots of the log-intensity ratios versus half log-intensity products for 6 of the 12 slides in this data sets. In the first 3 slides shown in Figure 2, hippocampus is labelled with Cy3(Green) and the remainder of the brain is labelled with Cy5(Red). In the second 3 slides, the dye scheme is reversed. The SL normalization curves are imposed on the M-A plots as green lines, the loess normalization curves are plotted as dotted red lines. Again, we see that they have similar overall trend, but there exist appreciable differences between them.

Tables 3 and 4 list the top 10 genes identified based on the proposed method and the method of Dudoit et al. (2000), respectively. The computation involved in Table 3 is similar to that in Table 1. The only difference is that for this data set, the number of clones is much fewer (557), the standard errors of the estimated expression differences are calculated directly based on the linear model theory without resorting to the bootstrap. The fact that the normalization curves are estimated from the data and its associated uncertainty are taken into account in this calculation. Again, the clones are listed in the order of the magnitude of their t-statistics. These two lists are completely different. Positive values of the t-statistics suggest that the clones are more highly expressed in brain excluding hippocampus, although the significance levels may not pass a desired threshold value. For the 10 genes listed in Table 3, the estimated mean differences are greater than those listed in Table 4. The standard error estimates are more stable and tend to be bigger than those in Table 4. The results of Table 3 correspond better to the results of sequence analysis of the cDNA libraries used in this experiment.

Figures 3(c) and 3(d) are similar to 3(a) and 3(b), respectively. The top 10 genes identified from the two methods are colored. In Figure 3(c), the red spots are the 10 genes from the SL method, and the green ones are the 10 genes from the loess method. However, for this data set, there are no common ones in the two top 10 lists. We see again that the bulk of the estimated expression differences from the two methods have the same distribution, but the values from the SL model tend to be higher. The correlation coefficient between the two estimated expression differences is 0.84.

5 Discussion

We proposed a SL model for normalization of microarray data and for identification of differentially expressed genes. It is interesting to compare the proposed normalization method

with the existing methods, such as the *loess* normalization proposed by Yang et al. (2001) and further discussed by Tseng et al. (2001). In the loess method, normalization is done separately by first fitting a separate curve for each slide through the scatter plot of y_{ij} versus $x_{ij}, j = 1, \ldots, J$ for $i = 1, \ldots, n$. Then the differences between y_{ij} and the value of the fitted curve at x_{ij} are used as the 'data' in the subsequent analysis. In comparison, we can write model (1) as:

$$y_{ij} - \phi_i(x_{ij}) = z_i'\beta_j + \varepsilon_{ij},$$

Thus in our proposed approach, the estimation of the gene expression levels is also done by use of the normalized values $y_{ij} - \phi_i(x_{ij})$.

However, there are three important differences between the proposed approach and that of Yang et al.(2001). First, the normalization curves $\boldsymbol{\phi}$ and the parameters of interest $\boldsymbol{\beta}$ are estimated simultaneously. With this integrated approach, there is no need to assume that the percentage of genes with differential expression levels is small, or when this assumption is not satisfied, to use dye-swap normalization, which in turn requires the assumption that the two normalization curves are symmetric. (However, we note that dye-swap as a design strategy is useful to balance the experimental conditions and reduce bias due to different dye incorporation efficiencies.)

Second, normalization for each slide in our proposed approach is dependent upon the data of all the sides in the same treatment group, with the $\boldsymbol{\beta}$ being the thread linking these slides. Each SL normalization curve does not attempt to fit data from an individual slide, it only fits the data from genes with constant expressions and those data for which the effects of differential expressions have been removed. In comparison, the loess normalization curve of Yang et al. only uses data from a single slide, and it is fitted from all the data from a single slide, which may lead to underestimation of the differences in gene expressions. Thus in general, the SL normalization can be more sensitive in detecting moderately differentially expressed genes.

Third, in the framework of the SL model, the uncertainty that is inherent in the normalization process can be taken into account in the estimation of the standard errors of $\boldsymbol{\beta}$. However, in the existing methods, such as that of Dudoit et al.(2000), the uncertainty from estimation of the normalization curves is not accounted for in the subsequent analysis.

The normalization we discussed can be considered as a type of location normalization, the purpose is to remove the potential bias due to imbalance between the two channels and other experimental factors. Sometimes it may be also necessary to perform scale normalization to make slides comparable in scale, as discussed in Yang et al. (2000), although scale incompatibility among arrays seems to be of a lesser problem in practice. We can extend the SL model to incorporate the scale normalization by introducing a slide-specific parameter $\boldsymbol{w} = (\tau_1, \ldots, \tau_n)$ as follows:

$$\frac{y_{ij} - \phi_i(x_{ij})}{\tau_i} = z_i'\beta_j + \varepsilon_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, J,$$

where τ_i's are restricted to be strictly positive. Computationally, we can proceed in a similar way as the algorithm described in Section 4.

In summary, the SL model provides a framework for combined normalization and analysis of microarray data. This method does not make the usual assumptions needed for the loess and dye-swap normalization procedures, nor does it require to identify a set of constantly expressed genes prior to normalization. It also naturally takes into account the uncertainty from the normalization process. For the two examples we considered in Section 4, the proposed method yield reasonable results when compared with the published results for the Apo AI data and the sequence analysis results for the hippocampus data. Thus the proposed SL model for microarray data is an interesting alternative to the existing normalization and analysis methods.

6 Acknowledgements

This research is supported in part by the NIMH grant K01-01541 [JH] and the U.S. Department of Energy grant # ER 62537 to MBS [JH, HCK, IL, and MBS].

References

1. Bonaldo MF, Lennon G and Soares MB (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Research 6, 791-806

2. Brown PO and Botstein D (1999). Exploring the new world of the genome with microarrays. Nat. Genet., 21 (suppl. 1), 33-37.

3. Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000) Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Genome Research, Vol. 10: 2022-2029.

4. Cleveland, WS (1979). Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc. 74, 829-836.

5. Drapper, N and Smith H (1980). *Applied Regression Analysis.* Wiley, New York.

6. Dudoit S, Yang YH, Callow MJ, Speed TP (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistical Sinica 12, 111-140.

7. Efron B (1979) Bootstrap methods: another look at the jackknife. Annals of Statistics 7: 1-26.

8. Efron B and Tibshirani R (1993) *An Introduction to the Bootstrap.* Chapman and Hall, London.

9. Engle RF, Granger CWJ, Rice J and Weiss A (1986) Semiparametric estimates of the relation between weather and electricity sales. Journal of the American Statistical Association, 81: 310-320.

10. Hastie T, Tibshirani R and Friedman J (2001) *The Elements of Statistical Learning.* Springer, New York.

11. Ihaka R and Gentleman R (1996) R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299-314, 1996.

12. Kerr and Churchill (2001) Experimental design for gene expression microarrays, Biostatistics, 2:183-201.

13. Schumaker L (1981) *Spline functions: Basic theory.* Wiley, New York.

14. Tseng GC, Oh M-K, Rohlin L, Liao JC and Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. Nucleic Acids Research, Vol 29, No. 12. 2549-2557

15. Westfall PH and Young SS (1993) *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment.* Wiley series in probability and mathematical statistics, New York.

16. Yang YH, Dudoit S, Luu P, and Speed TP (2001) Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Microarrays: Optical Technologies and Informatics, Vol. 4266 of Proceedings of SPIE.

Table 1: Top 16 genes from the analysis based on the SL model

ID	pvalue	t-statistic	numerator	denominator
2149*	0.0000	18.2949	3.3294	0.1820
540*	0.0000	18.2315	3.2283	0.1771
5356*	0.0000	11.5480	2.1336	0.1848
4941*	0.0000	6.4465	1.2130	0.1882
4139*	0.0000	6.3530	1.2481	0.1965
1496*	0.0000	6.3059	1.1445	0.1815
541	0.0000	5.4741	0.9822	0.1794
2537*	0.0000	5.4533	0.9721	0.1783
1739*	0.0000	5.3161	0.9599	0.1806
1337	0.0000	4.8553	0.9054	0.1865
563	0.0000	4.5950	0.8523	0.1850
3809	0.0000	−4.3686	−0.7641	0.1749
5986	0.0000	−4.2459	−0.8142	0.1918
4220	0.0000	−4.1181	−0.7748	0.1881
5722	0.0001	3.8561	0.7014	0.1819
947	0.0002	3.7764	0.7378	0.1954

Table 2: Top 16 genes from the method of Dudoit et al. (2000)

ID	pvalue	t-statistic	numerator	denominator
2149*	0.0000	21.5031	3.0806	0.1433
4139*	0.0000	13.6330	1.0251	0.0752
5356*	0.0000	11.6053	1.7957	0.1547
540*	0.0000	11.8907	2.9852	0.2511
1739*	0.0000	9.6767	0.8511	0.0879
2537*	0.0000	10.0097	0.9371	0.0936
1496*	0.0000	8.4200	0.9195	0.1092
4941*	0.0000	7.0476	0.9241	0.1311
947	0.0001	5.6995	0.6287	0.1103
5759	0.0002	5.0944	0.2196	0.0431
1932	0.0023	4.3768	0.2529	0.0578
4631	0.0013	−4.1695	−0.2292	0.0550
4160	0.0017	3.9402	0.2488	0.0631
5604	0.0018	3.9521	0.3661	0.0926
2324	0.0019	3.9362	0.3079	0.0782
926	0.0020	3.8513	0.3332	0.0865

The genes whose ID numbers have the superscript * appear in both Tables 1 and 2.

Table 3: Top 10 genes from the analysis based on the SL model

Gene ID	p-value	t-statistic	numerator	denominator
UI-M-BZ1-bkx-h-02-0-UI.s1	0.0000	-7.1099	-1.0537	0.1482
UI-M-AQ0-aae-h-02-0-UI.s1	0.0000	5.7338	1.0741	0.1873
UI-M-BZ1-bdp-f-01-0-UI.s1	0.0000	-5.5406	-0.8265	0.1492
UI-M-BH3-awc-g-02-0-UI.s4	0.0000	4.2744	0.6685	0.1564
UI-M-AQ0-aah-e-06-0-UI.s1	0.0001	3.9358	0.6428	0.1633
UI-M-BZ1-blk-g-12-0-UI.s1	0.0001	3.8087	0.5656	0.1485
UI-M-BZ1-bll-f-11-0-UI.s1	0.0003	-3.6384	-0.5387	0.1481
UI-M-AQ0-aaj-b-11-0-UI.s1	0.0005	3.5063	0.7772	0.2216
UI-M-AH0-acs-c-03-0-UI.s1	0.0005	3.4743	0.6459	0.1859
UI-M-BZ1-blc-e-02-0-UI.s1	0.0010	3.2816	0.4862	0.1482

Table 4: Top 10 genes from the method of Dudoit et al. (2000)

ID	p-value	t-statistic	numerator	denominator
UI-M-BZ1-bfv-c-02-0-UI.s1	0.0043	3.5756	0.2128	0.0595
UI-M-BZ1-bfw-g-12-0-UI.s1	0.0214	-2.6803	-0.1805	0.0673
UI-M-BZ1-bft-a-15-0-UI.s1	0.0232	2.6345	0.1759	0.0668
UI-M-BZ1-bdq-b-02-0-UI.s1	0.0243	-2.6091	-0.2186	0.0838
UI-M-BZ1-bfr-g-04-0-UI.s1	0.0254	2.5840	0.3028	0.1172
UI-M-BZ1-bjl-g-03-0-UI.s1	0.0343	-2.4156	-0.1577	0.0653
UI-M-BZ1-bli-e-04-0-UI.s1	0.0350	2.4040	0.1566	0.0651
UI-M-BZ1-bfs-e-08-0-UI.s1	0.0372	2.3692	0.1358	0.0573
UI-M-BZ1-beh-g-12-0-UI.s1	0.0406	2.3195	0.5970	0.2574
UI-M-BZ1-bjn-c-11-0-UI.s1	0.0413	2.3095	0.1684	0.0729

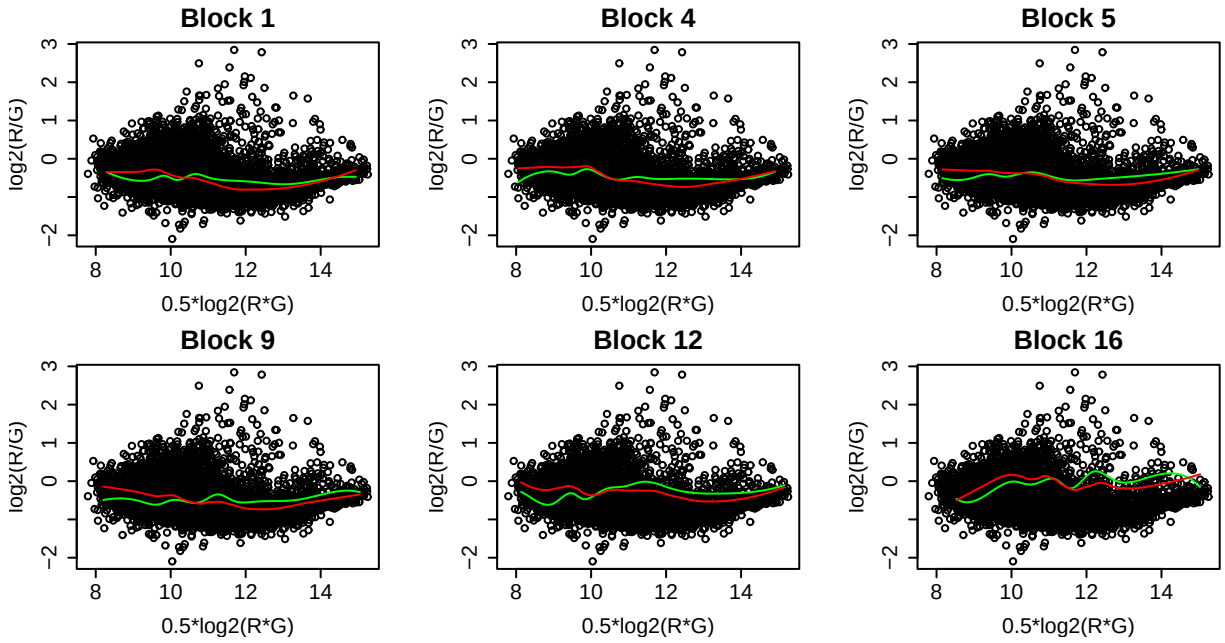

Figure 1: Apo AI data: Comparison of normalization curves in blocks 1, 4, 5, 9, 12 and 16 for the data from knock-out mouse 1 in the treatment group. Green line: normalization curve based on SPL model; Red line: normalization curve based on loess

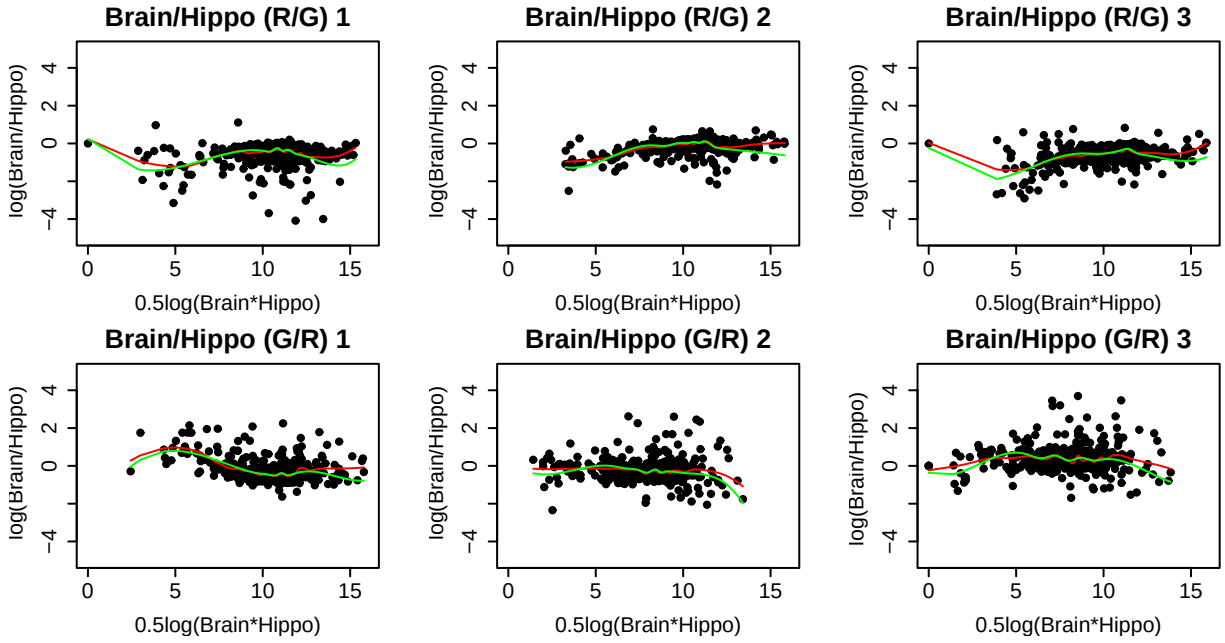

Figure 2: Hippocampus experiment data. Scatter plots of log intensity ratio versus log intensity product for 6 slides from the hippocampus experiment. In the first 3 slides, hippocampus is labelled with Cy3(G) and the remainder of the brain is labelled with Cy5(R). In the second 3 slides, the labelling scheme is reversed. Green line: normalization curve based on SL model; Red line: normalization curve based on loess

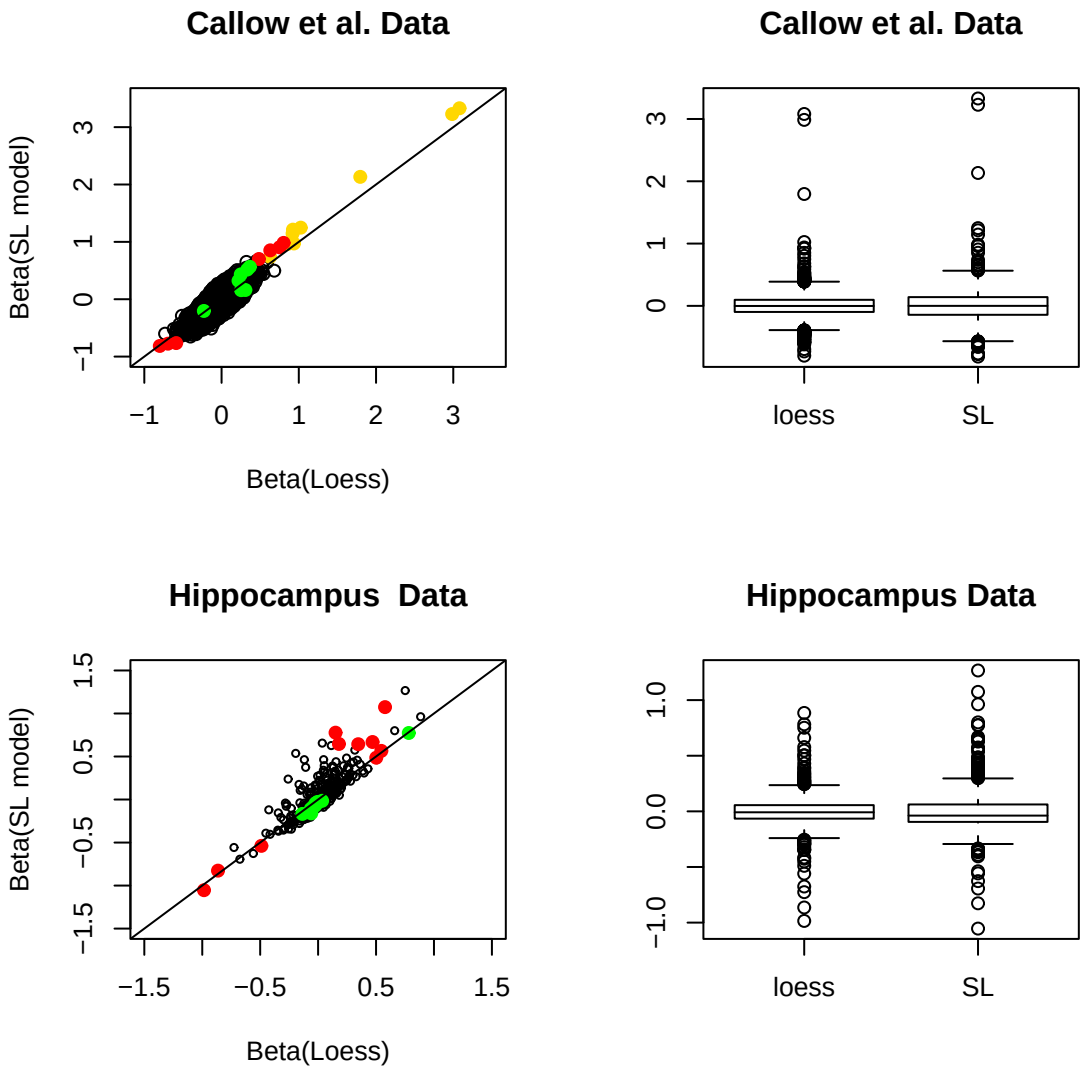

Figure 3: Comparison of estimated expression differences (EED) based on the SL method and the method of Dutoid et al. (2000). Figure 2(a). Apo AI data: scatter plot of EED based on the SL method and EED based on the method of Dutoid et al. Figure 2(b). Apo AI data: boxplots of EED based on the SL method and the method of Dutoid et al. Figure 2(c). Hippocampus data: scatter plot of EED based on the SL method and EED based on the method of Dutoid et al. Figure 2(d). Hippocampus data: boxplots of EED based on the SL method and the method of Dutoid et al.